DOCUMENT RESUME

ED 027 601 EA 001 974

By-Guba, Egon G.

Evaluation and Change in Education.

National Inst. for the Study of Educational Change, Bloomington, Ind.

Pub Date 16 May 68

Note-46p.; Paper prepared for the Elk Grove Training and Development Center Spring Evaluation Conference (Arlington Heights, Illinois, May 16, 1968).

EDRS Price MF-$0.25 HC-$2.40

Descriptors-Data Collection, Data Processing, Decision Making, *Educational Change, Educational Innovation, *Educational Planning, Educational Research, Educational Researchers, *Evaluation, *Evaluation Methods, Information Services

Identifiers-CIPP Model

Evaluation, defined as the process of providing and using information for making educational decisions, is a major factor in the achievement of planned change resulting in improved education. This kind of change is differentiated as neomobilistic or moving toward a planned new state, in contrast to evolutionary change and homeostatic change. Four distinct evaluations--of context, input, process, and product--are necessary in the neomobilistic change process. To perform the full range of evaluation tasks competently, a minimum of six specializations or evaluator roles is required--interpreter, instrument specialist, data processor, information specialist, reporter, and data collector. Six steps outline the procedure for designing an evaluation at each stage of the process. Ten criteria describe the requirements that must be met by evaluative information. (JK)

ED 027601

EVALUATION AND CHANGE IN EDUCATION

Egon G. Guba

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

A Paper Prepared for the
Elk Grove Training and
Development Center
Spring Evaluation Conference,
Arlington Heights, Illinois
May 16, 1968

EA 001 874

NISEC

The National Institute for the Study of Educational Change

825 EAST EIGHTH STREET • BLOOMINGTON, INDIANA 47401 • A.C. 812 337-9624

To judge by the number of requests that I have had recently to write about evaluation or about change, both of these topics are, in the parlance of the day, "hot." I have been fortunate enough to be able to relate to both areas, partly because of my interest in the Evaluation Center of the Ohio State University, which I helped to found before I left that institution and with which I have had a continuing relationship, and because in my current assignment at Indiana University I direct the National Institute for the Study of Educational Change. It is personally interesting to me that until I received the invitation to do this paper I had not really had the opportunity or the inclination to combine my thinking in these two areas, although it was patent that they were closely intertwined. Thus I relish the opportunity afforded to me by today's occasion and I hope that I will be able to make a useful response to it.

Obviously whatever one might say about the place of evaluation in the change process depends heavily on the conventions one adopts and the assumptions one makes. I feel compelled to begin my remarks, therefore, with some statement of what I shall mean by these terms. I hope that you will not find this introduction pedantic (although it probably will be) but I believe it is essential that we get straight in our minds just how these terms might be used and how they will be used by me. Having talked about evaluation and change as separate entities I will then attempt a formulation that puts them together. I will try to draw from this formulation some guidelines for the agencies and the individuals that might

be charged with certain of the functions that I will describe. I will also try to say something about procedures for designing evaluations, and determining their effectiveness.

It is customary in books and monographs to begin with a set of acknowledgements, but this practice is usually omitted in papers. It would be unconscionable of me not to mention some of the people on whose thinking I shall draw heavily, however, because the debt is so great. These persons include Daniel Stufflebeam, Director of the Ohio State University Evaluation Center and Robert Hammond, Director of Project EPIC in Tucson, Arizona, with whom I have had occasion to work intimately in recent months on ongoing evaluations; Walter Foley of the University of Iowa, William Gephart of Phi Delta Kappa, and Malcolm Provus of the Pittsburgh Public Schools, who are together with Stufflebeam, colleagues of mine on the national Phi Delta Kappa evaluation panel engaged in drafting a monograph on evaluation; John Horvat, my co-worker in the National Institute for the Study of Educational Change; Harbans Bhola, my former student, who taught me a great deal about the change process; and David L. Clark, my Dean at Indiana, with whom I have been able to collaborate over seven years in relation to issues in educational change. These eight persons will recognize many of their own thoughts here. I hope that they will not censure me too severely for my indiscriminate and often unrecognized borrowing.

The Nature of Change

Let me start by talking about the nature of change. The word "change" has many meanings, and we need to know something about these to be able to

draw certain analytic distinctions. Webster, my favorite author, lists among other definitions these three: (1) to become different, alter, vary; (2) to put or take a thing in place of something else, substitute; and (3) to cause to become different, alter, convert. These three definitions seem to impart different flavors to the term, representing what I shall call evolutionary, homeostatic, and neomobilistic change. I make these distinctions, which I shall define more carefully in a moment, because they are useful ones in relation to thinking about the change process and because they have different applications in relation to evaluation, as I shall try to show.

Evolutionary change is the kind that goes on all the time whether we know it or not. When we are in a cynical mood we may say that schools haven't changed much since we were children but that is in fact unlikely. It is even less likely that today's schools are essentially identical to those of the year 1900. It is certainly not the case that our schools are like those found in Puritan New England. Such change is mainly unconscious; it occurs over a relatively long time period, and its effects are difficult to detect by a contemporary observer even in a time of rapid change such as we are undergoing now. But it does occur, a fact which historians have no difficulty in documenting.

Homeostatic change may be either conscious or unconscious, but its essential feature is that it occurs only in reponse to some specific triggering. There are both good and bad ways to think about it. Harbans Bhola liked to refer to such change as reactive, implying by that term that it is mainly automatic and instinctively rather than thoughtfully guided. Horvat describes it as "reaction to disturbances in a system." The term homeostasis itself

has been adapted from the biological sciences, where it is used in a neutral sense to mean the mechanisms of system balance; when used in the behavioral sciences it is frequently given a pejorative sense, denoting defensive behavior. Obviously, however, homeostasis is also a good thing to have, as for example, to control body temperature or to restore blood pressure levels during a stressful period. I shall use it in this restoring or balancing sense, and shall try to show how evaluation can play an important role for that purpose.

Neomobilistic change is always conscious and always involves elements of planning or at least of intervention. The term neomobilism, by the way, cannot be found in your dictionary because it is one that Dan Stufflebeam and I made up one lazy afternoon when we had nothing better to do. We intend the term to convey the idea of change toward the new, or movement toward the new. Since neither of us are Greek scholars it is likely that we have committed some sins in concocting this word, but we think it catches up the essential idea. At least it has the advantage of not conjuring up in the hearer's mind other meanings that he has previously attached to the word.

The evaluation processes that relate most readily to homeostatic and neomobilistic change are quite similar and also quite different. We shall see that to some extent homeostiatic evaluation is concerned with detecting that which we know that we don't know, while neomobilistic evaluation is concerned with detecting that which we don't know that we don't know, if you follow me. The evaluation techniques are thus quite similar but the goals turn out to be somewhat different. But I shall return to this point in due course.

Neomobilistic change, and to some extent homeostatic change, at least when it is conscious, involve certain processes. First of all, there is the

identification of an area where change is needed. Sometimes this need is rather insistently thrust upon us, as for example, the problems of the disadvantaged youngster in the core city. Sometimes it is possible to identify an area that has been troublesome for some years without being dramatic in its impact. Data steadily mount up that indicate a problem exists and needs to be attended to, as in the case of gaining wider acceptance of vocational education. Sometimes we make a deliberate attempt to project what may still be some years in the future (as we are now doing for example through the new policy centers funded by USOE at Syracuse and Stanford). However it is done, whenever a need or a problem that require attention is identified, some kind of change is called for, ranging from a minor adjustment to a traumatic wrench.

Once a need or problem has been identified a response is called for. What can we do about it? Has anyone else ever done anything about it? Does anyone have any ideas about what might be done about it? Is there anything on the research shelf that might help? Does experience or precedent have anything to teach us? How can we tell whether any proposed response in fact meets the requirements posed by the need or problem? And if several possible responses meet those requirements, how can we determine which one is most likely to pay off? Thus, once we have determined that some change is needed, we need to find out what change should be attempted.

But having decided on some course of action is a far cry from having implemented it. To take a simple example, it is one thing to identify a reading problem resulting from the peculiarities of English spelling, another thing to see that some new kind of alphabet in which pronunciation is standardized might help with the problem, and a third and quite different thing to

produce the initial teaching alphabet (i/t/a/). This process calls for some kind of development, that is, developing a prototype of the idea, in practical form, separable from its inventor so that many people can use it, and tested to be sure that it responds to the need or solves the problem as required. The idea must thus be operationalized in usable form and packaged.

Despite the folklore to the contrary the man who builds a better mousetrap is not besieged by people beating a path to his door. No idea, even when put into easily usable form, ever achieved widespread use simply because, like Mt. Everest, it was there. Mt. Everest has a way of intruding itself on your consciousness--it is hard to ignore. But that is certainly not true of inventions. It pays to advertise in business, and it pays to diffuse in education. If you want people to use something they must first know about it. They must see that they have a need for it and they must understand that it will take care of this need better than what they have been doing or better than any other way in which they might be doing it.

Finally, even when a new solution to a problem or a new response to a need have been identified by a potential consumer, he is still faced with the task of integrating the innovation into his ongoing program. This task of institutionalizing may involve local trial, adjustments, reallocation of resources, administrative changes, and the like. Again a series of questions calling for decisions will emerge which must be handled properly if the innovation is to survive.

These steps that I have so briefly outlined constitute the normal progression through the change process. There are, it should be noted, many inputs,

of which research is only one and perhaps, in practice, a minor one. There are also many different ways in which the steps can be accomplished, and there are many different entry points into the process. Moreover, not all of the steps are equally possible for all types of educational agencies, and different contributions will be made by different groups. The process is very complex by its nature. I shall assert later that evaluation provides a means for ordering and rationalizing the process.

The Nature of Evaluation

Let me turn now to the second of the two aspects with which we are dealing, evaluation. Just as it is possible to define change in many different ways so is it also possible to define evaluation differentially. The question before us is not so much that of the "right" way to define evaluation, but is that of how we can recognize the contribution that different definitions can make to our thinking and how to devise a particular definition that suits the purposes and needs we have in mind.

In its earlier days educational evaluation was largely equated with measurement. This is not surprising in view of the history that preceded it. Evaluation really came into its own during the twenties following upon the heels of the very successful measurement movement. Abraham Kaplan is fond of what he calls the "Law of the Hammer," viz.,

> Give a small boy a hammer, and he will find that everything he encounters needs pounding.[1]

---

[1]Abraham Kaplan, The Conduct of Inquiry, San Francisco: Chandler Publishing Company, 1964, p.28.

Thus it was natural that following on the successful invention and adoption of standardized tests it was found that everything needed to be tested. The accumulation of scores, and the statistical manipulation of those scores to produce that pseudo-standard called norms, made it possible to form many judgments; this process came to be called evaluation. If you doubt that this definition still persists I invite you to open the catalog of any test publisher. In the introduction you are sure to be told that the collection of instruments therein offered for sale collectively adds up to an excellent evaluation program.

But this definition does have one important aspect, viz., it reminds us of the importance of being reliable, objective, and systematic about the way that we collect and interpret data. And that is a caveat to keep front and center regardless of your particular definition of evaluation.

Another definition which received wide currency, and is still the backbone of most evaluative thinking today, is that formulation which regards evaluation as a process for determining the congruence of performance and objectives. All school programs should be guided by behavioral objectives; indeed, it is the essence of program planning to project objectives and the essence of curricular planning to project a series of experiences through which the pupil can achieve the objectives. Similarly it is the essence of evaluation to determine whether the objectives were in fact met.

This definition has certain advantages and disadvantages. Clearly it is possible in these terms to focus not only on the student but also on the program. If the objectives are not met perhaps either is at fault; it is not a foregone conclusion that it is the student who is out of step. Thus feedback

is encouraged leading perhaps to diagnosis and remediation of students, but equally possibly leading to curricular change and refinement.

But at the same time evaluation is pointed by this definition toward outcomes; one cannot evaluate until performance has taken place and can therefore be compared to objectives. Thus while the definition implies feedback leading to refinement, such feedback typically cannot occur until the termination of the treatment being evaluated. The promise of the definition is thus often not attained.

Moreover, the matter of criteria remains unresolved. While standards are perhaps implicit in the statement of objectives the source of the objectives is mystical. It is often asserted that they are "screened" through a philosophy and a psychology but which philosophy and which psychology remains unresolved. So too does the meaning of the term "screen." Finally, this definition of evaluation places an overwhelming importance on behavioral objectives; we have come to believe that only ultimate evaluations which trace everything back to improved student performance are worthy of the name. Today we insist on evaluating even national programs like ESEA Title III this way, forgetting that there are other standards that might be applied with greater validity. Thus it would not occur to us to evaluate a carburetor by looking for changes in driver behavior, but in education an analogous process seems to make sense.

Another definitional problem has to do with the relation of evaluation to research. In some minds these terms are equivalent, but even when they are differentiated, it is usually assumed that the proper methodology of evaluation is the methodology of research. Evaluations are thus designed the way

experiments are designed. Evaluative data are collected and interpreted the way research data are collected and interpreted. And evaluative conclusions are research conclusions.

No doubt this equating has a great deal of utility. It affirms a great deal of technique as appropriate and invokes certain scientific standards that allow us to judge the value of the work. But it also has enormous disadvantages. Research is concerned with establishing universally true new knowledge, applicable to a well defined population. Its requirements are such that controls must be carefully observed. Research tells us what happens in the best of all possible worlds, as for example, when a feather and a stone are simultaneously dropped in a vacuum. But evaluation is none of these. It is not concerned with universal knowledge but particular knowledge. It is concerned with what works. It tells us what we may expect in the real world, as when a feather and a stone are dropped on a windy day in Central Park. It invites interference, rather than controls it, because it is concerned with such real world questions.

Most importantly of all, research methods are inappropriate to evaluation because they pose constraints that prevent successive improvement and refinement, which is the essence of planned change. Given a research viewpoint, the treatment to be evaluated must be specified at the outset. Once the "start" button is pressed nothing can be changed. No other unusual things (e.g., the introduction of a second new program) can be permitted during the evaluation period. Data are not available until the termination of the experimental period, too late to do anything about what has happened. Finally, unless possible sources of problems were foreseen and built into the design there will be no data to tell much about why what happened happened. If a non-significant difference is found

who can say why? Process data just are not available. And without them the treatment cannot be improved. The essence of research method, control, which is intended to make the findings context-free (i.e., universally applicable) thus is exactly wrong for evaluation which is very much context-oriented and developmental in its nature.

A definition of evaluation which obviates some (although not all) of the difficulties I have enumerated and which has certain utility in relation to the change process is currently under formulation in a number of centers of activity. This definition views evaluation as a process of providing and using information for making educational decisions. I do not have the time today to indicate all of the thinking that went into this particular, deceptively simple formulation. I will assert without providing argument or evidence that this definition sees evaluation as continuing (cyclical and looping), multifaceted (involving many different methods and techniques), practical, and relevant. Its obvious relation to change is that the definition allows us to view evaluation as a tool for dealing with the many different kinds of decisions posed in my recital of the change process earlier in this paper. Indeed, evaluation can be set off into a number of types or categories as a function of the kinds of decisions to be served. Let me turn to that problem now.

Types of Evaluation

If evaluation is to serve the decision-maker it is obvious that different types of evaluation may emerge if we can postulate different types of decisions. Indeed, it would be most useful from a practical point of view if it were possible to develop a complete taxonomy of educational decisions; we might then be able to

describe all possible evaluation situations. While some of us are working on this problem it is by no means solved, and hence we must make do with apparently useful approximations. One of the most useful formulations, in my judgment, has been proposed by Stufflebeam[2] in the so-called CIPP Model (CIPP is an acronym formed from the names of the four basic types of evaluation that are identified). Stufflebeam classifies educational decisions into four categories, i.e.,:

> . . . planning, programming, implementing, and recycling. Planning decisions are those which focus needed improvements by specifying the domain, major goals, and specific objectives to be served. Programming decisions specify procedure, personnel, facilities, budget, and time requirements for implementing planned activities. Implementing decisions are those in directing program activities. And, recycling decisions include terminating, continuing, evolving, or drastically modifying activities.[3]

Corresponding to these four classes of decisions are four kinds of evaluation that service them, which are called context, input, process, and product evaluation respectively.

Context evaluation, as the name implies, is mainly concerned with providing information about the context or setting within which the educational activity is taking place. It includes that kind of information which nowadays is often called baseline data. The term baseline data, however, implies a

---

[2]For a full explication of Stufflebeam's point of view see his paper "Evaluation as Enlightenment for Decision-Making," an address delivered at the working conference on assessment theory sponsored by the ASCD Commission on Assessment of Educational Outcomes, Sarasota, Florida, January, 1968. The paper is available from the Ohio State University Evaluation Center, Columbus.

[3]Ibid., p. 30.

static description of a starting point, as for example, where a class stood in relation to certain achievement variables at the time when a new program was introduced. Context data, as that term is used here, go well beyond this static concept. They are not collected once but continuously, and their nature may be constantly changing as new requirements and needs are introduced. A good context evaluation provides status data whenever and as often as may be required.

But a context evaluation does more than merely to describe what is going on, whether at one point in time or at multiple points. First of all, it begins by describing the limits of the domain to be considered, and it determines, within those limits, what sub-areas or topics are most crucial to know about and what data will yield the desired information. Thus context evaluation properly applied is emergent, i.e., it is self-determining and does not just conform to some pre-determined design. Further, context evaluation not only describes status but gives indications of when and how that status is other than what it should be; in this sense context evaluation is a kind of process control or monitoring device that indicates when a decision needs to be made. It does not respond just to decisions that can be forecast but it also forecasts decisions that should be made and which it will subsequently serve. Thus a context evaluation should not be viewed as a data bank or a social book-keeping operation, both of which can be pre-planned. Context evaluation is self-generating or self-programming to a major degree.

It may be useful to examine the likely products of a context evaluation. There will of course be descriptive data, and these descriptive data will be

more encompassing, more flexible, and more timely than we are accustomed to expect. But the data will be organized to indicate needed decision situations; thus the decision-maker will receive not only reports but interpretations that tell him that he needs to do something. Moreover, the thing about which he needs to do something will be described in terms of specifications that suggest reasonable responses. The specifications not only provide concreta to which the problem solver may respond but also provide criteria in terms of which any proposed solution or response may be judged. The results of a context evaluation will be similar to the specifications which an architect develops for a house: they contain certain descriptive data (e.g., a family of five is involved with three being youngsters); they contain certain requirements (e.g., space is needed to store two cars); and they contain certain criteria (e.g., the house must allow a traffic pattern that militates against crossing the living room to get to the kitchen). The specifications may also raise certain decision situations that had not earlier been recognized, e.g., whether or not the water faucets shall be single or double action. Information of this kind both illuminates the decisions to be made and exposes the variables that must be considered in making them.

Input evaluation has to do with decision alternatives, their delineation, and their relative worth. It provides the information necessary to program a response to the need or problem which is the focus of the decision. And, it should be noted, it is carried out in terms of the variables and criteria that have already been defined through the context evaluation.

When the architect has developed the specifications for a house he is able to project a number of plans that meet those specifications more or less

well. The object of the input evaluation game is to make some assessment of whether a reasonable variety of possible plans have in fact been proposed, and to determine, from among the available plans, which one meets the criteria best or has the highest payoff potential in that situation. Thus, the various plans for a home may be evaluated in relative terms for the space provided, the cost per square foot, the traffic patterns afforded, etc.

The problem of criteria is a very difficult one in making these assessments. Usually the decision situation is not a simple one, i.e., limited to a single decision and a single criterion. If the architect is to build that house which gives the maximum square-footage to cost ratio, he has a simple task: build a loft. But at the same time he has to meet certain requirements of taste, of structural soundness, of fittingness to the community in which the house is to be built, and the like. Optimizing one criterion may very well minimize some other criterion. Moreover, the decisions may turn out to be contingent, and that what appeared to be reasonable at one level of decision making turns out to be very troublesome at another. Thus, the decision to build a two story house because the ratio of area to cost is lower for such a building produces problems in designing an air conditioning system that will maintain regular temperature since the second floor will normally be warmer through convection currents. Finally, it is often the case that there are multiple decision-makers invoking multiple criteria; thus a husband and wife may not agree on whether a basement is needed.

Despite these difficulties it is fairly easy to specify what the products of an input evaluation should be: a reasonable set of alternative responses, arrived at either by compiling those in existence or by generating new ones by

an ad hoc method, and a weighting of those alternatives, in probabilistic form, to indicate their success potential. The decision-maker is thus confronted with a range of possible choices and some basis for deciding among them.

Process evaluation is related to implementation of the selected alternative. It is concerned with signaling process decisions that need to be made, and providing data that illuminate those decisions. Process evaluation is, to use an analogy, the guidance system, both in the sense that it signals whenever the program is off course and also in the sense that it provides a mechanism for correction.

The term "debug" is especially descriptive of one of the intents of process evaluation. Any new element inserted into an ongoing system can go wrong in a variety of ways. It is likely that not all of the contingencies were foreseen by the planners, but they will become quickly apparent in operation. Some means is required to detect them as soon as possible and to determine their dimensions so that appropriate corrective action can be taken. If, for example, the teacher is expected to use an oral-aural rather than a grammatical approach in teaching language, it is important to have some check on whether she actually is doing that, and if not, to institute corrective action. Or if it is anticipated that certain materials will be available about Thanksgiving time for distribution to the class, we need to know whether they arrived or not. All the possible misunderstandings that can arise in communication, all the backslidings into more familiar or comfortable modes of doing things, all the logistical diversions, all the attitudinal interventions, all the physical factors, all the administrative arrangements, that

are required must be checked to be certain either that things are going as planned or to pinpoint the source of difficulty.

But the process is concerned with more than debugging. It also has as a major objective further refinement and extension of whatever is being evaluated. The development and perfection of an innovation intended to meet some need or to solve some problem is a complicated process requiring many attempts. Hopefully each successive effort will be a better approximation of what is needed. Process evaluation is concerned with identifying in what particulars further refinements are needed, i.e., with helping to decide how the approximation can be made to fit better. In this sense process evaluation is rather like systems or operations research, in which the system or operation being developed is tested, refined, tested again, and so on, all the time that the system is in operation. In designing a better hospital organization, for example, one cannot suspend the hospital while the new organization or procedure is being developed; rather, it must be installed into the ongoing system and then perfected. It is with this perfecting, as well as with debugging, that process evaluation is concerned.

Product evaluation is concerned with final outcomes. In this sense product evaluation is rather similar to the conventional mode of defining evaluation in terms of congruence between performance and expectations. Given this perspective it is possible to see why the traditional mode of defining evaluation was characterized as narrow earlier in this paper; within the emergent definition conventional evaluation is only one of four possible modes.

The major decision with which product evaluation is concerned is whether or not the item being evaluated should be retained, eliminated, or altered in some way. Hence Stufflebeam has characterized this mode of evaluation as particularly suited to recycling decisions.

The major data generated by product evaluations relate to the effectiveness and efficiency of the treatment being evaluated. Here it is entirely appropriate to ask whether the objectives have in fact been attained. We must understand, however, that the term objectives need not mean pupil behavioral objectives, however. These may or may not be reasonable objectives; surely they will not be if the focus of the evaluation is outside the scope of classroom learning.

An Example

This brief discussion summarizes what is meant by context, input, process, and product evaluations. An illustrative example of the entire sequence may be found in the situation surrounding the so-called Cuban missile crisis. Let us examine that briefly.

The CIA by means of its daily U-2 Cuban overflights noted that certain objects, apparently missiles, were present in Cuba. Intelligence experts were called in to examine the photographs, in an effort to identify the objects positively and to make some assessment of their power, range, scope, and the like. These experts determined that the missiles posed an impossible security situation for the U.S. mainland in terms of certain standards that had previously been determined by defense authorities. Obviously something had to be done. The situation was presented to the decision-maker,

the President, who now knew that he had a decision to make. He was also provided the data necessary to determine the domain of the decision, its criticality, and the important factors or variables to be considered in making the decision, (power, striking force, and the like). The context evaluation was then complete.

At this point the President surrounded himself with a group of advisors who had the capability both to generate alternative responses to the crisis and to assess the impact and utility of each. While we are certainly not privy to what actually occurred at this time, we can speculate that a number of possible retaliatory actions were considered, such as bombing Cuba, invading Cuba, serving an ultimatum on Russia demanding withdrawal of the missiles, and of course, quarantining the island. We may presume that each was discussed from several points of view. Would that action really remove the threat to U.S. security? Would the problem be solved temporarily or permanently? What were the probabilities that any retaliation would touch off World War III? What would happen to the morale of our civilians? How would our allies respond? And so on. Finally after due consideration one alternative emerged, not as a sure-fire solution, but as that response which, on balance, would produce at least a minimally effective response without incurring too high a probability of undesirable complications. Thus, the input evaluation was complete.

Now we had an action decision, but it remained to be implemented. First there was the question of whether an effective quarantine could be mounted. Could orders be drafted for the ship commanders that would strike the delicate balance between firmness in the pursuit of our goal and deference to the rights of other foreign powers? Would the orders be understood? Could the commanders be counted on to carry them out? Could the

ships be moved into their positions in time? Could all sea lanes be closed effectively? There were also the questions of possible feedback and refinement. How could one capitalize on the early experiences in stopping and searching to make the blockade more effective? What about the patrol patterns of the vessels--was there some method that would reduce the likelihood of penetration? When intelligence procedures had been set up so that information of this sort could be collected and used to check on and improve mounting and maintaining the quarantine, then process evaluation would have been served. It should be noted, however, that this evaluation needed to be continuous to be maximally effective.

Finally, of course, there was the question of effectiveness. This question could be answered proximally, e.g., in terms of whether the blockade was effective in intercepting all vessels in Cuban waters and determining that there were no missiles or other contraband aboard. There was also the more ultimate question of whether or not the missiles already in Cuba could be forced out by this procedure. As it turned out, the quarantine was indeed effective, requiring a minimal interaction with the enemy to achieve its purpose. Thus product evaluation was also accomplished. But, it might be noted, the way in which it was finally ascertained that the missiles were removed was to revert to the same evaluative methods that had percipitated this entire decision and evaluation process--the U-2 overflights. We see then that product evaluation finally becomes context evaluation again, and that the whole process can by recycled. In this case we may hope that the continuing context evaluation will raise no more red flags.

Evaluation and Change

Having told you a great deal more than you really wanted to hear about evaluation and change as separate entities, let me turn now to the question of their relationship. I trust that my discussion has been lucid enough so that you will already have guessed how I will link these two processes.

First, let me dispense with evolutionary change as having little relationship to evaluation, or perhaps I should say, that evaluation has little relation to it. Evolutionary change occurs so slowly that it is not subject to the ordinary measurement procedures of the behavioral sciences; probably only the methods of the historian or the anthropologist have any relevance. In any event, evolutionary change is not beset with decisions, since it occurs unconsciously and without guidance. Evaluation as defined here cannot serve any useful purpose.

But homeostatic and neomobilistic change can be served very well indeed by evaluation, and I will try to describe an evaluation approach which is appropriate to each.

In relation to homeostatic change, I will assume that I am discussing a situation in which homeostatic change occurs by design and for purposes of institutional maintenance and adjustment, not for reactionary reasons relating to maintenance of the status quo. I assume that there is a decision-maker who is concerned with detecting problems relating to the organization or agency on whose behalf he makes decisions, and that he requires some method for detecting such problems and responding to them rationally. I will further assume that most of the decisions he will need to make can be

predicted in advance, but that there will occur some situations in which he is, so to speak, taken by surprise. Nevertheless he will need to be able to respond to these situations as well.

The prudent decision-maker in such a situation will provide himself with the analog of the U-2 overflights, that is, with a way through which he can monitor his agency or organization and find out whether everything is going as it should. Now the glib terms "everything" and "should" must be defined, of course. What is everything and how can it be monitored systematically? In a school, for example, do I monitor student achievement only, or staff morale, cafeteria cleanliness, bus scheduling, and community attitudes as well? And when I do monitor whatever I monitor, how can I tell when things are going as they should, i.e., what criteria shall I invoke? If I find 5 per cent student underachievement do I alert a decision-maker? What if two teachers are found to have low morale? And so on. We shall return later to the matter of how these determinations might be made.

Now let us suppose that we have such a system, i.e., a context evaluation system, in operation. Note, incidentally, that the system is collecting data continuously and must be prepared to respond to the responsible decision-maker whenever he has scheduled such a response or whenever a response seems to be required by some anomaly in the context data. Now insofar as the system reports "A-OK" the decision-maker rests easy. But if the data indicate some deviation beyond previously defined limits, or if the data indicate that some problem has come up which requires attention, the decision-maker must take action. We may call these deviations <u>non-congruences</u>, and hence we may refer to homeostatic evaluation as <u>congruence</u> evaluation.

What the decision-maker must do is to respond in a way that will correct the deviation or solve the problem, leaving everything in a state that will cause the evaluation again to report "A-OK." He may make this response automatically, drawing upon previous experience with similar situations in which he has found a certain response to be highly effective. Or he may search out the alternative responses and assess them for probable effectiveness. He is then engaging in input evaluation.

The homeostatic input evaluation is likely to be automatic in many cases, perhaps even in most cases. But at times the problem may require some special inputs that have not been available or even invented before. Thus when it is discovered that the existing curriculum just is not successful in teaching certain kinds of children to read, the superintendent may appoint a special curriculum committee charged to develop some new way of reaching these children. Or when the district has become so geographically dispersed that bus schedules fail to work and children frequently are late arriving at school, a contract may be let with a local computer agency to devise a computerized approach to the scheduling problem. It seems safe to say that whenever a response to a need or problem identified through context evaluation cannot be made more or less automatically an ad hoc input evaluation will be required. The input evaluators are likely to be not professional evaluators but substantive experts who can propose new solutions and suggest ways in which their probable success can be estimated. The professional evaluator can take over at that point to collect the data that may be required actually to project success potential.

The input evaluators now report back to the decision-maker who looks at what they have done. He may accept their suggestions or he may recycle the whole operation, as for example, when the solution with the highest projected success potential still has only a minimal chance of succeeding in absolute terms. But let us suppose that the proposed "best" alternative is acceptable. That alternative must now go through the engineering and testing process that was described in my earlier remarks on change. Here process and product evaluation comes into play. These evaluations are also ad hoc, being devised to fit the particular innovation under consideration. These evaluations are probably carried out by professional evaluators, since we are here talking mainly about the technical determination of whether an innovation is meeting its goals and doing it efficiently.

When the findings of the process and product evaluations are presented to the decision-maker he is in a position to decide what to do with the innovation. He may determine that the whole thing looks so bad that recycling into a further input evaluation is required, to generate a different approach. He may decide that the innovation, while promising, requires further refinement, and he may call for a recycling through another process and product evaluation. Or he may decide that the innovation is "proven" well enough to warrant permanent introduction to the operating system.

If he makes this latter decision the decision-maker must provide for diffusion and institutionalization. The generation of an appropriate diffusion strategy may also require evaluations of the sort described here: a context evaluation to determine such matters as the nature of the audience to whom diffusion is to occur, the assumptions concerning the end state in which the

audience is to be left, the assumptions concerning the nature of the innovation as it relates to the audience (e.g., is it a mere substitution for this group or does it go deep into their values?), and so on--the elements, in short, that help to determine a suitable strategy. An input evaluation may devise alternate strategies and weight these for probable success. The actual diffusion program can undergo a process and product evaluation, and so on. Thus evaluation does not end by any means at the point at which an innovation is ready for distribution, but has great utility for shaping the character and thrust of the diffusion effort.

Finally, when the innovation is introduced to the system, alterations must be made in the context evaluation system so that the innovator, too, can be continuously monitored along with other important operational elements. Whatever was done in process and product phases can be adapted for this purpose, although data collected for context purposes will probably not be as extensive. Nevertheless the important thing to note is that we have moved from the ad hoc status of input, process, and product evaluations back to the normal and systematic context evaluation, and once again the decision-maker rests easy.

The discussion thus illustrates several points about homeostatic, or, as I prefer to think of it, congruence evaluation:

1. There are three separate loops in the total evaluation system, one of which, the context loop, is continuous and persistent, while the other two, input and process-product, are ad hoc.

2. The context loop is set up to detect whether the system is operating within previously described limits and/or whether some unpredicted problem has arisen which requires attention.

3. The context data provide a dynamic baseline which process and product evaluations can use as a reference point. Similarly, changes in context evaluation systems are made by reference to process and product data.

4. The input loop is activated only when a non-congruence situation arises to which a response cannot be automatically made. The loop can be recycled as often as necessary to produce an apparently viable response.

5. The process-product loop is activated only when an innovation has emerged from the input loop which is to be tested. This loop can also be recycled as often as necessary. Its data may call for an even more complex recycling back to the input loop.

6. When an innovation has been "proved" it is plugged into the context loop and the whole process recycles.

7. The end result of congruence evaluation is homeostasis. Any alterations or changes which it produces tend to restore the system to equilibrium by overcoming disturbances.

Let me turn now to the matter of neomobilistic evaluation. If homeostatic or congruence evaluation has as its intention the maintenance of equilibrium, neomobilistic evaluation has as its intention moving the system out of equilibrium, into a new arena, and into equilibrium there. This form of evaluation closely parallels Kurt Lewin's three-phase change process model of unfreeze, move, and refreeze.[4] Unfreezing is related to a determination of the fact that some movement is desirable or necessary, moving is related

---

[4]Kurt Lewin and Paul Grabbe, "Conduct, Knowledge, and Acceptance of New Values," Journal of Social Issues, 1 (1954), 64.

to getting the system to the new place, and refreezing is related to forming and maintaining a new equilibrium. The difference between homeostatic and neomobilistic evaluation is rather like the difference between the process control unit of industry and the research and development unit. One is concerned with maintenance while the other is concerned with new directions. And obviously no industry could long exist if it were lacking either kind of mechanism.

The major difference between homeostatic or congruence evaluation, on the one hand, and neomobilistic, or, as I shall call it contingency evaluation, on the other, is in the context evaluation loop. In the homeostatic case the context is given; it is internal to the system which is being evaluated. Congruence evaluation is, as I have pointed out, analogous to the process control mechanism typically found in industry, and its domain is the domain of the organization involved. Process control is concerned only with what is already ongoing, and its concern is to be sure that the ongoing lies within given limits and is not threatened by any unexpected problems.

Contingency evaluation, on the other hand, is external to the system, and hence its proper domain might be anything that is conceivably related to the operation. But what is conceivably related is not easy to determine; the identification of this domain depends heavily on the creative imagination of the persons involved in the evaluation. There are no inherent boundaries. Thus contingency evaluation is rather analogous to the research and development units found in most industries; their domain is whatever they take it to be.

What is it that such a unit should monitor, and what kind of data should it collect? Obviously, to carry the industry analog further, it can

monitor the "competition," that is, other agencies and mechanisms that are engaged in essentially the same kinds of operations. Such monitoring is useful both in the sense of helping solve the problem of what data to collect (what would be appropriate to the new ideas that other people are following up on?) and of providing red flags that indicate that what the parent organization is doing no longer is keeping up with what others are doing.

But the contingency context evaluation team needs to do more than just stay in touch with other agencies--it needs to be creative about what might be done by the parent organization. Thus we have here an element of forecasting the future; indeed, of planning the future so that the general objectives of the parent organization may be extended and refined in new arenas. The policy centers recently funded by USOE at Syracuse and Stanford are examples of context evaluation agencies set up at the highest national level with the mandate of at least predicting the future (I suppose that a federal agency cannot, overtly at least, declare itself in the business of planning the future in our culture). But future planning is by no means an idle or impossible game. Shane and Shane have recently discussed this concept as follows:

> Future-planning is proposed as a procedure for creating curricular and instructional strategies that are more than hindsight remedies for today's problems. It employs a sophisticated means for combining values as well as data from education and related disciplines. These, together with the power of controlled imagination, are deliberately employed to create the particular future that our beliefs recommend from among the many less desirable futures in which education, by default, may find itself.[5]

---

[5]Harold G. Shane and June Grant Shane, "Future-Planning and the Curriculum," Phi Delta Kappan, 49 (1968), p. 372.

It is clear that if the intent of this statement is followed up, that future-planning is concerned not merely with predicting what the future might be (indeed, the cynics of today would suggest that such a prediction would be dour indeed) but also with actually creating that future.

Once the context evaluation effort has exposed a series of contingencies that might call for a redirection of effort, the decision-maker to whose attention these contingencies are called is faced with an input evaluation situation. He now assembles a group competent to generate alternatives and to weight them for success potential just as he did in the case of congruence or homeostatic evaluation. The major difference between input evaluation in the two situations is not so much a matter of technique as it is in the competence of the persons engaged to perform this task. They need to be more creative, more visionary, more change oriented, and certainly more willing to take risks. But then the risks should be higher for the stakes are higher too--the results may change a whole segment of the real world rather than simply to adjust one portion of it.

Process and product evaluation differ little in this neomobilistic situation ~~form~~ FROM the homeostatic situation. There is some innovation to be tried, and essentially the same techniques work in both cases. The innovation resulting from the neomobilistic approach may be larger in scope and perhaps more pervasive in its consequences but the evaluation problems differ little.

From this discussion we may extract several points about contingency evaluation:

1. Like congruence evaluation contingency evaluation consists of three loops: context, input, and process-product.

2. These contingency loops resemble the congruence loops in the ways in which they are activated and in the ways in which they can be recycled.

3. The major difference between the context loops in congruence and contingency evaluation is that the domain of the former is largely given whereas the domain of the latter is undefined and continuously changeable.

4. Contingency evaluation is less sure of its direction than is congruence evaluation. The risks of low return for its activities are higher.

5. Contingency evaluation has the capability not only to predict the future but also to help to create it.

Agency Responsibility for Evaluative Tasks

I have led you a merry chase through a maze of definitions and it may be worth while to stop a moment to catch our breath. Let me remind you briefly at least of the major terms that I have introduced. First, I defined evaluation as the process of providing and using information for decision making. Then I asserted that there were four kinds of evaluation corresponding to four broad classes of decisions which educators needed to make, and I defined these as context, input, process, and product evaluation. The first of these, I said, was continuous, while the other three were combined into two loops that were activated as the need arose. Finally, I said, there were two major types of evaluation, homeostatic and mobilistic (these two I also called by the names congruence and contingency evaluation respectively--there is no difference). I indicated that these terms were

analogous to process control and research and development counterparts in industry.

Both congruence and contingency evaluators keep up a continuous monitoring. So long as the results of this monitoring spell "A-OK," as it were, no further action is necessary. But when some response is called for, either in terms of future-planning or in terms of deviation correction, input loops are activated to generate alternative responses and to weight these for success potential. The selected alternative is then engineered (not by an evaluation group) and put through the process-product loop to determine its overall utility. Once the tested innovation is installed and institutionalized all loops except context are deactivated. No further action is then necessary until the context loop again indicates a need.

The question I want to raise with you now is this: given that you go along with everything that I have suggested so far, just what agencies are to carry on all of these various kinds of evaluations? Should every local school district, for example, be staffed and budgeted to carry out context, input, process, and product evaluations? Who does congruence evaluations and who does contingency evaluations?

A variety of conditions and factors enter into such decisions. These include personnel availability, funding levels, support base, and operational needs, among others.

It seems clear that every agency ought to have some kind of context evaluation capability--indeed, most agencies now already have such a mechanism in at least rudimentary, perhaps even informal, form. All that is implied

by this statement is that all agencies ought to have ways of keeping track of what is going on. The bookkeeper in the superintendent's office performs such a function, for example, as does the person who counts the number of cafeteria lunches sold or the number of IQ tests administered that year. All normal accounting and social bookkeeping functions are included. Perhaps the only real difference between keeping track of what goes on in most schools and what I am suggesting here is in the rigor, thoroughness, and continuity of the data collection and interpretation process, in the definiteness with which the decisions to be served are identified a priori (so that they guide the data collection operation), and in the extent to which criteria are formally identified. But these are technical, albeit difficult, matters.

The main difficulty in mounting an ongoing context evaluation seems to be a matter of economics. Some school districts simply are not big enough or wealthy enough to be able to afford a data collection and interpretation system that will do the job efficiently. Here some kind of cooperative effort with other districts may be called for. My own guess is that a base of about 20,000 pupils is needed.

When one moves to the level of input, process, and product evaluation, we are talking about another matter, however. Here we are not concerned simply with operational data that flow on a more or less continuous basis. The work level may be sporadic and the needs, as we have already noted, are indeterminate. The response must therefore be ad hoc. It follows that a larger base is needed. Moreover, many of the problems identified in school district A will parallel those in school district B. The expense and overlap

involved in having both districts work on the sample problem is probably unwarranted, particularly from the taxpayer's point of view. I would suggest therefore that a new agency be established, which Dan Stufflebeam and I have been calling a "halfway house" in our own conversations, that would perform at least the input, process, and product evaluations necessary for a number of cooperating districts, and perhaps, for smaller districts, the context evaluations as well. I would guess that a pupil base of at least 100,000 would make sense as a guideline. These halfway houses could of course be funded nicely under the provisions of ESEA Title III. They could include funds for the expensive computer and other machines that would be useful for such a center to have at its disposal. They must also include, of course, the development teams required to take any proposed alternative that has been turned up through input evaluation and engineer it into a viable and useful invention.

When one turns to the question of congruence vs. contingency evaluation, we are clearly talking about different <u>levels</u> of operation. No doubt most local school districts will be concerned primarily with congruence evaluation. Their problems will be mainly operational problems; in any event they will not have the staffing nor other necessary capabilities to do broad scale future planning. New agencies are necessary for this task. Only recently, as I have already mentioned, the Office of Education funded two such planning centers at Syracuse and Stanford. Unfortunately because of the possible stigmata of federal control and big brotherism attached to any federally funded project of this sort, it will probably not be possible for these agencies to do more than future

prediction. There is here a broad field, I believe, for private and foundation capital to support genuine future planning efforts.

These two centers are only a beginning, however. Moreover, such projects need not all be at the federal level; there will be plenty of opportunity at regional, state, and even sub-state levels to engage in such contingency evaluation. But new agencies will be necessary. I hope that it will be possible for every state department of education to establish such a unit, perhaps using ESEA Title V money. The data collected would be of inestimable value not only to the unit to which they would accrue but to every other level including the local level.

Individual Evaluator Roles

Let me turn now to the problem of personnel. It is apparent from the preceding discussion, again assuming that you are placing some credence in what I have to say, that the term "evaluator" has received a considerable broadening. Indeed, there are a number of specializations that seem to emerge which should be pointed out, since not all of these specialists are now being trained, or if trained, may perhaps be receiving inappropriate orientation.

Let me begin with this last observation first. Most persons who would label themselves, or be labeled by others, as specialists in evaluation have been trained as educational researchers. I have tried to indicate in one or two places the specific problems produced by this approach from a technical point of view, but the problem is broader than that.

Mainly, it seems to me, we are dealing here with questions of values or orientation. Not only are the research-trained evaluator's techniques inappropriate but he persists in approaching every problem as though it were a research problem, and as though his objectives were to establish universally true laws. As a result he typically feels that the persons bringing the evaluation problem to him don't know what questions to ask, and he feels that he must reinterpret their problems into a broader scientific framework. He usually ends up asking the questions that interest him rather than the questions that need to be asked. Moreover he approaches his relationships with the evaluation client in a rather condescending way, which, in addition to being professionally diverting, also produce an infuriated client. I contend that until we recognize evaluation as a scientific field in its right, with its own appropriate values and methodology, we shall find ourselves terribly constrained and unproductive.

If we were training persons specifically for the tasks that I had outlined, what kinds of personnel would we want to produce? It seems to me that we need to think in terms of six rather different roles:

1. The evaluator as interpreter. If evaluation is intended to provide information for decision-making it follows that the evaluator must be in close and continuous contact with the decision-maker. This involves an intimate professional relationship which provides perhaps the evaluator's chief claim to fulfilling a genuine professional role. In this situation the evaluator needs to work with the decision-maker, whom we shall refer to as his client, in order to determine (a) the nature of the decisions to be

served, (b) the variables which the client believes are important in making this decision, and (c) the nature of the standards or norms which the decision-maker will apply in interpreting the relevant data which may be provided. Once these facts are known the evaluator can serve in the role of team leader, directing other members of the evaluation team in their tasks. Finally, the interpreter will take the developed data and cast them into a form relevant and focussed for the client's needs.

That this role will be a most delicate one to fill should be apparent. The evaluator in this role can get by with little technical proficiency but he needs a great deal of human relations proficiency. He must be skilled at reading between the lines and making professional inferences. We must remember that he will be working with a decision-maker who will have little time to give him, and that the decision-maker will probably have little insight into the ways he makes decisions or the standards he invokes. He may even pride himself on his ability to "shoot from the hip," and may be stunned to find how inconsistent or irrational his traditional mode of operation is. He may resent the evaluator who reflects these shortcomings back to him. He may be unable or unwilling to indicate some of the "real" standards that guide his actions, for example, political considerations. But the evaluator's task is not to critique his standards but to service them; he cannot castigate the client for invoking political considerations (indeed what person attuned to the real world would not?) but must do his best to collect accurate data about what the political situation <u>is</u>. I might point out that it is at precisely this point that today's research oriented evaluator often forsakes his client, on the grounds that

standards such as political considerations are irrelevant and inappropriate. Indeed they are, in the case of scientific inquiry, but certainly not in the case of evaluation.

Eventually the evaluator must also report back to the client. His information must be timely and it must be credible. If the client does not trust the evaluator and his methods he will surely place little credence in any of his information but particularly in his sensitive information, such as political analysis. In that case the evaluative information is worse than no information at all. It has taken time and resources to collect, it is wasted, and it may actually serve to reduce the decision-maker's propensity to utilize evaluation effectively in the future.

Thus this interpretive role of the evaluator is crucial, both in the sense of defining what information is necessary, focussing the tasks to be accomplished to generate that information, and reporting that information to the decision-maker in credible and timely form. The evaluator as interpreter works mainly in interaction with clients and needs to have the appropriate skills. Comparatively speaking it matters little whether he can program a computer or run a chi-square.

2. The evaluator as instrument specialist. Whatever information is necessary for the client must be collected, and means must be found to collect it. In some areas instruments already exist, for example, to measure student achievement (although even here the validity of the instruments may be questionable). In most cases I would say, instruments will have to be constructed. I use the word "instruments" very broadly to indicate all data collection devices, paper and pencil tests, interview schedules, observational schedules, etc. included. The construction of valid and reliable instruments for these purposes requires someone with special skills.

This role is of course more technical than that of interpreter. The instrument constructor may never see the client, and probably will rarely see the subjects from whom his instruments will extract data, except perhaps during periods when the instruments are themselves being tested. Generally speaking, this person does not need human relations skills so much as he needs broad technical proficiency in devising ways to collect information that may be needed, whatever its form or sources.

3. The evaluator as data processor. The information collected in any evaluation comes in what is often called "raw" form, which strikes me as implying that it should somehow be "cooked" before it becomes palatable. The cook in this case is the data processor. Again we are talking here about a technical role. The data processor must be conversant not only with statistics in order to be able to run significance tests (insofar as these may be useful) but also with computer operation so that he can devise efficient means of managing the data flow and to produce the required outputs. Little client relations skill is necessary, although it might be useful to know something about the personality quirks of computers (!).

4. The evaluator as information specialist. A quite different and I believe new evaluator role called for by the present rationale is that of information specialist. I do not refer here so much to an information specialist in the storage and retrieval sense, although such aspects may be involved, but to an information specialist in the substantive sense called for by input evaluation needs. When a problem is identified we need to know what, if anything, already exists that might be useful in relating to

that problem, whether it be research, experience, solutions invented elsewhere, or whatever. The information specialist is charged with devising means of collecting that kind of input data.

Now presumably there exist mechanisms through which such information might possibly be available, for example, the ERIC system of USOE. But the existing systems, I believe, are only automated ways of obtaining what one could always obtain with a few research assistants and a good library at one's disposal. The trouble with such information sources is that they yield abstracts or listings of information, not interpretations. What is needed for input purposes (and indeed, for the later development and packaging tasks associated with engineering the proposed solution into a viable package) are not abstracts but <u>interpretations, modules of information</u> that do not presume substantive expertise on the part of the ultimate consumer. A series of abstracts indicating, for example, what sociologists know about the core city does me little good unless I am trained to interpret and use sociological data. And no one can be trained in all of the areas which might have something of relevance to contribute. Hence the task of the information specialist is not just to dredge up the information but somehow to formulate it into pre-interpreted modules that are of use to the client. It is this latter task that at once differentiates this information role from that of the storage and retrieval expert and simultaneously raises its professional level a good many degrees. I do not know how to train such a person nor do I know of any institution which is now trying to develop an appropriate training program. Yet the need is very evident to me.

5. The evaluator as reporter. The wealth of information which might be dredged up in relation to some evaluative problem is probably too great to transmit in toto to the decision-maker. There is no point in telling him more than he needs to know; to go beyond those needs is to waste his time. What is required therefore is a reporting specialist who can cast the information into easily communicated form and who can focus it in ways designed to facilitate the decisions which will be made. This is not simply a reportorial or editing role, it should be noted, but one that requires a good knowledge of the decision requirements and the ability accurately and objectively to select and present relevant data without introducing biases. Again I know of no agency now training such persons.

6. The evaluator as data collector. This last role for which we must train persons seems to me to be more para-professional than professional but its importance must not be minimized. Ultimately in every evaluation empirical data of some kind must be collected, and probably from subjects in field situations. Someone must supervise the data collection and perhaps even carry it on himself, depending upon the complexities involved. So for example participant observers, interviewers, classroom observers, or test proctors may be needed. Any evaluation group will have to have persons who can structure the required field conditions, train persons on an ad hoc basis to collect data, and supervise them in that collection. Universities have not produced such people--possibly because their roles are not "professional" enough to warrant advanced degrees (or to make it possible even to grant advanced

degrees given the typical university's graduate program requirements and admission standards), but clearly we need to move in this direction.

These six roles represent the minimum amount of professional differentiation required to perform today's evaluation tasks competently. In general we are not training such persons now, at least in a differentiated way. Moreover, as we do train and employ such people we will undoubtedly uncover many operational problems involved in getting such a team to function well together. For example, since most of these roles will be filled by highly trained specialists, perhaps with Ph.D. degrees, we will have the problem of giving visible professional credit to each. An evaluative report cannot simply bear the name of the team leader, say. Each specialist may need to be given wide latitude to participate in activities beyond those defined for his role, so that he can feel that he is making a broad contribution. The problem of advancement along clear career lines must also be attended to. I have no answers now to these questions but merely point them out as requiring attention.

Designing an Evaluation

The person charged with actually operationalizing an evaluation is likely at this point in my remarks to be wondering how in heaven's name one can ever get anything as ponderous as what I have been describing functioning effectively. How can one design an evaluation that would meet all of these requirements?

To deal with this problem in detail would require more time than I have at my disposal. I would like to point out that there are systematic

steps for getting this job done, however. An excellent guide for designing an evaluation has been developed by Stufflebeam and his associates which at least outlines the key questions and problems that must be faced. He indicates that the following steps are involved regardless of whether the evaluation being designed is a context, input, process, or product evaluation:[6]

1. Focusing the evaluation. In this step, carried out by the interpreter-evaluator, we must identify the levels of decision-making involved, the decision situation to be served, the criteria to be invoked in reaching the decision, and the general policies within which the evaluation must operate (for example, the degree of publicness).

2. Collection of information. Here we must, for each decision situation, specify the source of the information to be collected, the instruments and methods to be employed, the sampling procedure involved, and the conditions and schedule for information collection.

3. Organizing the information. This step must result in the specification of a format for the collected information and a means for coding, organizing, storing, and retrieving it.

4. Analyzing the information. The analytic procedures to be employed and the means for performing the required data analysis must be specified.

5. Reporting the information. First we must be concerned with defining the audiences for the several evaluation reports. Then, means for providing required information to each audience must be specified, and the format for presenting the information to each audience must be developed. Finally a suitable reporting schedule must be devised.

---

[6]See Stufflebeam, op. cit., p. 41ff.

6. Administering the evaluation. In this step, which permeates all of the other steps, we must define staff and resource requirements and plans for meeting those requirements, evaluate the proposed design in terms of appropriate criteria (which I will briefly mention in the concluding section), specify the means for periodic updating of the evaluation design, and of course, project the necessary budget.

These six steps, which can be operationally specified in much greater detail, provide a rough framework to guide the evaluator as he designs the evaluation. They are difficult, time-consuming, and they require frequent and continuing interaction with the decision-maker(s) to be served.

Evaluating the Evaluation

I should like to turn very briefly now to the matter of how the evaluator can evaluate his own activity. The information which the evaluation produces is the key. What criteria are appropriate to it?

This question can be answered in two parts. If evaluation produces information, then that information must meet criteria that are ordinarily required of any good information. But because it is evaluative information it must also meet certain special criteria.

The general criteria are these:

1. Internal validity. The information provided by the evaluation must display a reasonable correspondence to the phenomena which it purports to describe or interpret. It must have fidelity, or, in the layman's sense, it must be true.

2. External validity. The information must be generalizable to similar situations beyond the one in which it was collected. Particularistic data have little utility. If for example data relating to the effectiveness

of an innovation could not be interpreted as also holding in classrooms other than the ones in which they were collected we would gain little in deciding whether to adopt or not.

3. Reliability. Here we are concerned with the replicability of the data. If a repetition of the evaluation did not produce essentially similar findings we should be concerned that the findings were simply random and therefore meaningless.

4. Objectivity. Here our concern is with the publicness of the data. If they are private in the sense that only particular persons would so interpret them, i.e., that not all competent judges would agree on them, we can only wonder about their true meaning.

In addition to these four general criteria that could be invoked in relation to any information, certain special criteria must be met by evaluative information. These are:

5. Relevance. The information must relate to the decisions to be made.

6. Significance. The information must be weighted for its meaning in relation to the decision. Not all relevant information is equally weighty. The culling and highlighting required is a professional task that justifies the inclusion of a reportorial expert on the evaluation team.

7. Scope. The information must relate to all aspects involved in the decision. If there are six alternatives to be considered information that applies to only four lacks in scope.

8. Credibility. The information must be trusted by the decision-maker.

9. Timeliness. The information must come in time to be useful to the decision-maker. The evaluator must guard against the scientific value that argues against publishing findings until every last element is in. Late information is worthless information. It is better in the evaluative situation to have reasonably good information on time than perfect information too late.

10. Pervasiveness. The information must get to all of the audiences (i.e., to all of the decision-makers) who need it.

An evaluator who can say, after careful examination, that his evaluation design will produce information that conforms to all of these criteria can be assured that he is doing his job well.

Finale

I have undoubtedly already tried your patience beyond all bounds. It is likely that I have been a bad evaluator in judging what information you wanted or needed and in reporting it to you in relevant terms. I have probably overstressed evaluation and talked too little about change. But I do not believe that anyone who sees the intimate connection that I see between evaluation and change as processes would believe that. Indeed, if it were not an obvious overstatement, I might say, if only for effect, that evaluation and change are in many ways indentical processes. Instead, however, I will simply say that if, as is often asserted, mathematics is the queen of the sciences because all others are dependent on it and are serviced by it, that evaluation is certainly the queen of the change processes, for similar reasons. I can only hope that I made this point well enough that you will seek to pursue it on your own.